文書ベクトル表現を用いたネットワークによるニュース記事推薦手法の提案

今井 貴之 †　　中村 啓太 ‡　　大豆生田 利章 †
† 群馬工業高等専門学校 生産システム工学専攻　‡ 会津大学 復興支援センター
ap15803@ipc.gunma-ct.ac.jp, keita-n@u-aizu.ac.jp, mame@ice.gunma-ct.ac.jp

1　はじめに

インターネットの普及により，誰もが情報の提供・配信を容易に行うことが可能になった．現在，Web 上には膨大なデジタルコンテンツが存在し，その量は日々増加傾向にある．これに伴い，ユーザが閲覧するデジタルコンテンツも増加しており，ユーザの望む情報かどうか判断するために大変な時間と労力を必要としてきている．特に，デジタル文書は日々莫大な量が配信されており，ユーザはこの中から必要な情報を探すため，一つ一つの内容を確認する必要がある．このような負担を減らすため，ユーザの嗜好に合わせた情報推薦システムや支援システムの研究 [1][2] は多く行われている．

本研究ではデジタル文書の中でもウェブサイト上のニュース記事を対象とした情報推薦を目的とする．ニュース記事を内容に基づいて分類し，ネットワークで可視化し，推薦を行う手法を提案する．ニュース記事に対し，BoW(Bag-of-Words) によりニュース記事を文書ベクトルに変換する．これらの文書ベクトルのコサイン類似度を算出し，これを用いてネットワークを生成する．これによって生成されたネットワーク内のサブネットワークに対し，次数中心性を用いてそのサブネットワークを表す代表的なノードを抽出する．抽出された各ノードの文書ベクトルに LDA（Latent Dirichlet Allocation）を適用する．LDA を適用したこれらのノードの文書ベクトルに対して，コサイン類似度を算出し，ネットワークを生成，可視化を行う．これにより情報を集約し，また集約したノード量によってノードの大きさを変えるなど，ネットワークの特性を活かした可視化を行うことで，効率の良い情報推薦システムの構築を目指す．

2　関連研究

Web 上のコンテンツを推薦することを目的とした研究は多く存在する．多数のエンドユーザが Web ページにタグ情報を付与することで分類を行う Folksonomy マイニングに基づく推薦システムを構築した研究 [3] では，Folksonomy を利用することで，ユーザの嗜好情報が解析でき，タグの表記ゆれに対策を講じることで効率的な情報推薦を行っている．また，Blog 記事を収集し解析することによって得られる Blogger の嗜好を利用した強調フィルタリング方式の Web 記事推薦システムの開発を行った研究 [4] では，Blog の利用者数の多さを利用することで，強調フィルタリングの課題であるコールドスタート問題の発生を抑制した推薦システムの開発を行っている．しかしながら，これらの研究では速報性を有するニュース記事に対して，タグが十分に付与される，または Blog 記事が更新されるまでのタイムラグは非常に大きなものとなってしまうため，ニュース記事に対する効率的な推薦手法とはいえない．

さらに，Twitter のフォローユーザに重要度を設定し，それらの Tweet 情報を利用することで，ユーザの関心があるニュース記事を推薦する研究 [5] では，ユーザの嗜好に合わせたニュース記事を推薦することを可能にしている．この手法ではフォローユーザの数が膨大な場合，情報過多になってしまう，重要度設定が困難になるなど，ユーザ数によって生じる問題が挙げられる．

本研究ではこれらと異なり，内容に基づくことで新規のニュース記事を確実に分類し，ネットワークを用いた可視化によって効率化した推薦手法を提案する．

3　提案手法

本研究で対象としているニュース記事は日々増加しており，その量は膨大である．そのため，これらを一

度に分類すると，分類後も情報量が多くなってしまう．そこで本手法では，内容が近似しているニュース記事をまとめ，一つのニュース記事とみなすことで，情報の集約を行う．また，集約されたニュース記事の関係性をネットワークで示すことで，ユーザが閲覧しているニュース記事に関したニュース記事の推薦を行う．以下では，文書分類手法，ネットワークによるニュース記事集約手法およびネットワーク可視化手法について説明する．

3.1 文書分類手法

文書を内容に基づいて分類するため，本研究ではBoW を用いてニュース記事を単語による文書ベクトルに変換し，コサイン類似度によってニュース記事間の類似度を算出する．

BoW とは文書を単語の集合として考えるモデルである．ある文章にどの単語が出現したかのみを考え，並び方は考慮しない．ニュース記事に対し，形態素解析機 MeCab[6] を使用し，文章を単語に分けて，「名詞」の単語のみを抽出する．抽出された名詞の単語群を BoW で用いる集合とする．文書ベクトルに変換した後，それらの類似度をコサイン類似度を用いて算出する．

式 (1) にコサイン類似度を示す．ここで，A, B は文書 A,B の文書ベクトルとする．

$$sim(A, B) = \frac{A \cdot B}{|A||B|} \tag{1}$$

以上の方法によって算出された類似度を用いて，ネットワークを生成する．

3.2 ネットワークによるニュース記事集約手法

ネットワークとはノードと呼ばれる点とリンクと呼ばれる線で構成されるグラフである．本手法ではノードをニュース記事とし，類似しているノード間をリンクでつなぐことでネットワークを生成する．生成されたネットワークはいくつかの小さいネットワークに分かれており，これらをサブネットワークと呼ぶ．このサブネットワークに対し，次数中心性を求めることで，サブネットワーク内で最も代表的なノードを抽出することができる．

次数中心性とはネットワーク理論で扱われる中心性の指標の一種である．中心性とは，ネットワーク内で最も中心的なノードを示すものである．次数中心性はノードの中で最も次数が多いものが中心的なノードであると考える手法である．これによって同一サブネットワーク内のノードを一つの代表的なノードに集約することができる．

3.3 ネットワーク可視化手法

前節の手法により，抽出されたノードの文書ベクトルに LDA を適用し，単語による文書ベクトルをトピックによる文書ベクトルに変換する．これに対してコサイン類似度を用いて類似度を算出し，ネットワークを生成，可視化する．

LDA[7] とは文書の生成過程を確率的にモデル化したトピックモデルの一種であり，一つの文書には複数のトピックが混合されていると仮定し，各単語ごとの潜在的トピックを決定するモデルである．これによって，単語ごとにトピックが決定され，BoW で用いられた単語次元のベクトルをトピック次元のベクトルに圧縮することが可能となる．本手法では Collapsed Gibbs sampler を用いた LDA[8] を構成している．これにより，トピックによる類似の幅が生まれ，より推薦向きな分類を行うことができる．

生成されたネットワークでは一つのノードに集約されたノード数によって大きさを変化させたり，時系列を色別で示したりすることで，可視化による情報推薦の効率化を図る．

4 比較実験

複数のニュースサイトで掲載された記事に対し，提案手法を用いて生成したネットワークと既存手法（TF 値と N-gram を用いた手法）[9][10] で生成したネットワークを比較及び検討する．

4.1 実験内容

本研究では Yahoo!ニュース，読売新聞，朝日新聞，日経新聞，産経新聞のサイトにおいてトップ記事として掲載されたものをニュース記事として使用する．これらの記事に対して提案手法を用いてネットワークを生成，可視化を行う．提案手法において，本実験ではネットワークによるニュース記事集約手法におけるリンクを接続する条件をコサイン類似度が 0.65 以上の数値があるものとする．また，ネットワーク可視化手法においてはコサイン類似度が 0.25 以上であればリンクを接続する．可視化されたネットワークにおいて，ノードの大きさはニュース記事集約手法で集約されたノードの量に応じて変化させる．また，ノードの色はニュース記事の取得日時別に色分けを行う．既存手法でのしきい値は 0.1 とする．本実験では 2016 年 11 月 22 日から 2016 年 11 月 28 日までの期間で収集した計 2919 件

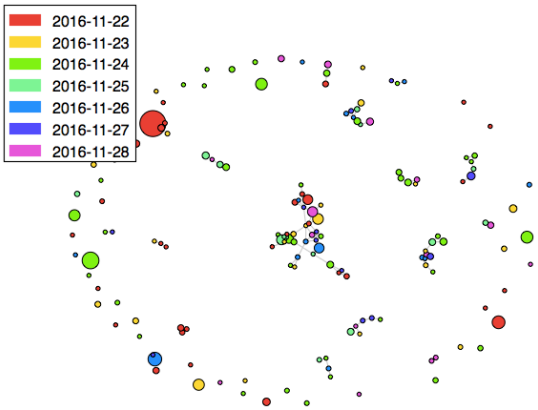

図 1　提案手法によるネットワーク全体図（ノード数 155，リンク数 282）

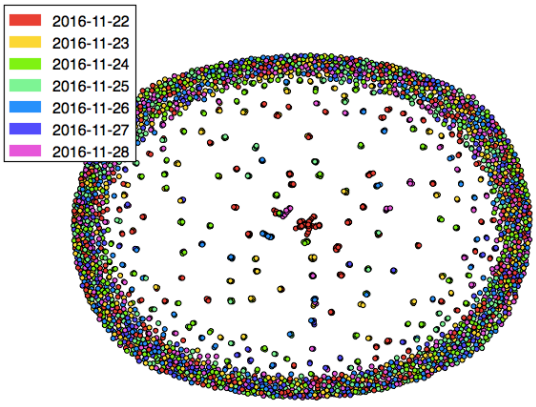

図 2　既存手法によるネットワーク全体図（ノード数 2919，リンク数 1301）

のニュース記事を扱う．

4.2　実験結果・考察

図 1 に提案手法によって生成したノード数 155，リンク数 282 のネットワークの全体図を示す．また，図 2 に既存手法によって生成したノード数 2919，リンク数 1301 のネットワークの全体図を示す．これらの図より，提案手法のネットワークは既存手法のネットワークに比べて，ノードを集約することで情報が整理されていることがわかる．さらに，図 1 のネットワークではあるニュースについての記事がどれだけ多く投稿されたかによってノードの大きさが変わるため，ニュースの注目度を視覚的に把握することが可能であるとわかる．

次に，それぞれの図 1,2 において 11 月 22 日に発生した福島県沖の地震についてのニュース記事のサブネットワークについて比較した結果を示す．図 3 に提案手法による福島県沖の地震についてのサブネットワークを示す．このサブネットワークのノード数は 4，リンク数は 3 である．また，表 1 に図 3 のそれぞれのノードに集約されているノード数を示す．次に，既存手法による福島県沖の地震についてのサブネットワークの一部を図 4，5 に示す．図 4 のノード数は 59，リンク数は 285 であり，図 5 のノード数は 7，リンク数は 12 である．図 3 と図 4，5 を比較すると，図 3 のサブネットワークの方がより情報を集約できており，さらにサブネットワークも一つにまとまっていることで，ネットワークの特性を活かした関連記事の推薦を行うことが可能である．また，図 4 のサブネットワークに存在するノードは，図 3 における記事番号 0 番のノード内に集約されているため，より推薦に優れたサブネットワークが提案手法によって生成できたといえる．

以上より，既存手法と比べて，提案手法によって生成したネットワークの方が情報推薦において有効的であるとわかる．

5　おわりに

本研究では BoW を適用して，ニュース記事を文書ベクトルに変換し，これらから算出したコサイン類似度を用いてネットワークを生成した．生成されたネットワーク内のサブネットワークを次数中心性を用いて一つのノードとみなすことで，情報の集約を行った．抽出されたノードに対して LDA を適用し，それぞれのノードの文書ベクトルを単語次元からトピック次元に圧縮した．これらに対してコサイン類似度を計算し，それらを用いてネットワークを生成，可視化を行った．さらに，既存手法によって生成したネットワークと比較実験を行い，結果として，提案手法は既存手法よりも多くの情報を集約し，ネットワークの特性を活かすことのできるネットワークを生成することができた．

今後の課題として，二つのことが挙げられる．

一つ目に，分類精度についてである．今回の実験では経験則に基づいたしきい値を用いたが，この数値が推薦を行うにあたって適切であるかどうかの評価実験を行う予定である．さらに，ユーザの嗜好に合わせてしきい値を変更するなど，ユーザ情報を用いた推薦手法も検討中である．

二つ目に，推薦手法の評価についてである．本研究の最終的な目的は，内容に基づいたニュース記事の推

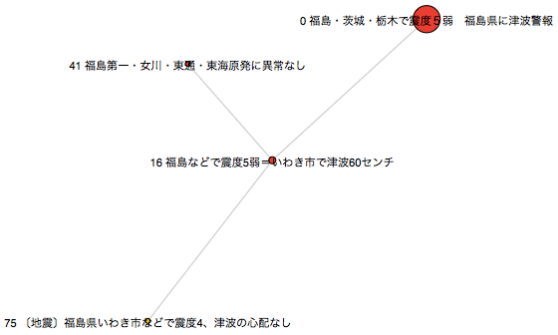

図 3　提案手法における福島県沖地震についてのサブネットワーク（ノード数 4，リンク数 3）

表 1　提案手法における福島県沖についてのノード概要と集約ノード数

記事番号	ニュース記事概要	ノード数
0	福島・茨城・栃木で震度５弱	120
16	いわき市で津波 60 センチ	8
41	福島第一原発に異常なし	4
75	いわき市などで震度 4	4

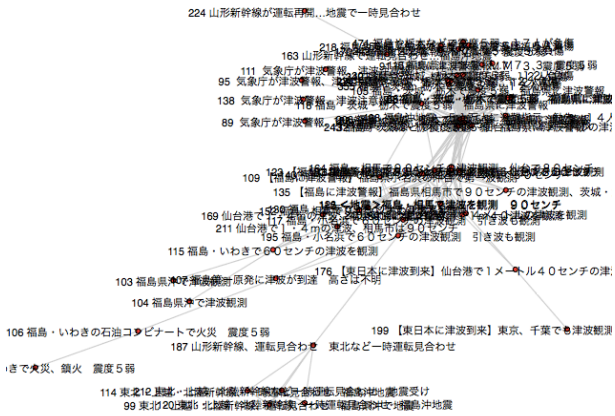

図 4　既存手法における福島県沖地震についてのサブネットワーク 1（ノード数 59，リンク数 285）

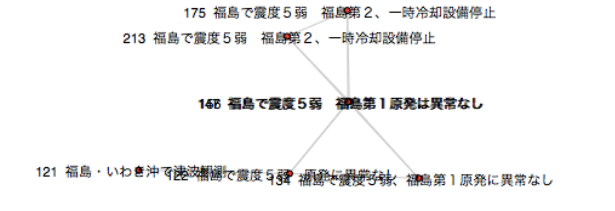

図 5　既存手法における福島県沖地震についてのサブネットワーク 2（ノード数 7，リンク数 12）

薦システムの構築である．そのため，本手法を推薦システムに導入し，ユーザに効率的な推薦を行えているか，評価実験を行う予定である．

参考文献

[1] 渡邊恵太, and 加藤昇平. "トピックモデルと協調フィルタリングに基づくユーザ興味を反映した情報推薦システム." 2014 年度人工知能学会全国大会, 2M3-4 (2014).

[2] 平田紀史, et al. "時系列を考慮した階層的クラスタリングに基づくインタラクティブなニュース記事閲覧支援システム." *The 23rd Annual Conference of the Japanese Society for Artificial Intelligence*. 2009.

[3] 丹羽智史, et al. "Folksonomy マイニングに基づく Web ページ推薦システム." 情報処理学会論文誌 47.5 (2006): 1382-1392.

[4] 小原恭介, et al. "blogger の嗜好を利用した協調フィルタリングによる Web 情報推薦システム." 人工知能学会全国大会論文集 0 (2005): 133-133.

[5] 早川豪, et al. "Twitter を利用したソーシャルニュース記事推薦システム." 研究報告データベースシステム (DBS) 2011.16 (2011): 1-4.

[6] MeCab: Yet Another Part-of-Speech and Morphological Analyzer: `http://taku910.github.io/mecab/`, (2017/1/15 アクセス).

[7] Blei, David M. et al. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

[8] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." *Proceedings of the National academy of Sciences* 101.suppl 1 (2004): 5228-5235.

[9] 今井貴之, et al. "ネットワークを用いたテキストマイニングによる類似ニュース記事の可視化." 言語処理学会第 22 回年次大会 (2016) : 913-916.

[10] Takayuki Imai, et al. "Visualization of Similar News Articles with Network Analysis and Text Mining." *2015 IEEE 4th Global Conference on Consumer Electronics* (2015): 151-152.